A COMPANION ESSAY · NAIROBI · 2026

Measuring What Actually Matters

Kirkpatrick levels 3 and 4: how you find out whether any of it changed behaviour or helped a patient.

Dr Neal Aggarwal

Medicine · Surgery · Engineering · Information Technology · Artificial Intelligence

17 AUGUST 2026

The tenth of the Institute's ten pedagogical commitments is one sentence long: *we measure at Kirkpatrick 3 and 4, or we admit we do not know.* **This essay unpacks that sentence completely**, starting from zero: what the four Kirkpatrick levels are and why almost every training programme stops at the second one; what "behaviour" means when the behaviour in question is a habit of mind; how you would actually observe, audit and log a student-clinician at three and twelve months without fooling yourself; why the independent-impression rule is the fastest-decaying thing taught, and what would refute that; what a stepped-wedge design buys you and costs; and a working catalogue of the other pedagogical instruments — entrustment, programmatic assessment, Angoff, retrospective pre-post, logic models, audit and feedback.

Contents

The post on one sheet: the Kirkpatrick staircase from satisfaction to clinical results, the independent-impression rule and the three-month booster that Level 3 exists to check, the eight-encounter minimum that makes a behavioural measure reliable, and the Level 4 discipline — pre-registered indicators, balancing measures, and a stepped-wedge trial, an interrupted time series or a logic model depending on how much of the rollout you control. Click for the full-resolution version.

Watch this first — a quick summary of the text that follows. Eight minutes on the whole argument: why a satisfaction score at the end of a course tells you almost nothing, what the four Kirkpatrick levels actually ask, why Level 3 — whether a clinician does it on the ward, months later, when nobody is watching — is the one that costs money and the one that gets dropped, and what it would take to measure it honestly. Everything below is the same case at length — the instruments, the decay curve, the study designs and the honest limits — so the video is the shape of it and the text is the substance.

Download the video

Listen to this post (45 minutes): A deep dive into commitment ten — why happy sheets are close to worthless and what they are still good for, how Levels 1 and 2 are built for student-clinicians, what Level 3 looks like in operational detail (mini-CEX, chart audit, the independent-impression rule, the three-month booster), how a Level 4 result can be claimed without fooling yourself, and what this whole apparatus still cannot tell you.

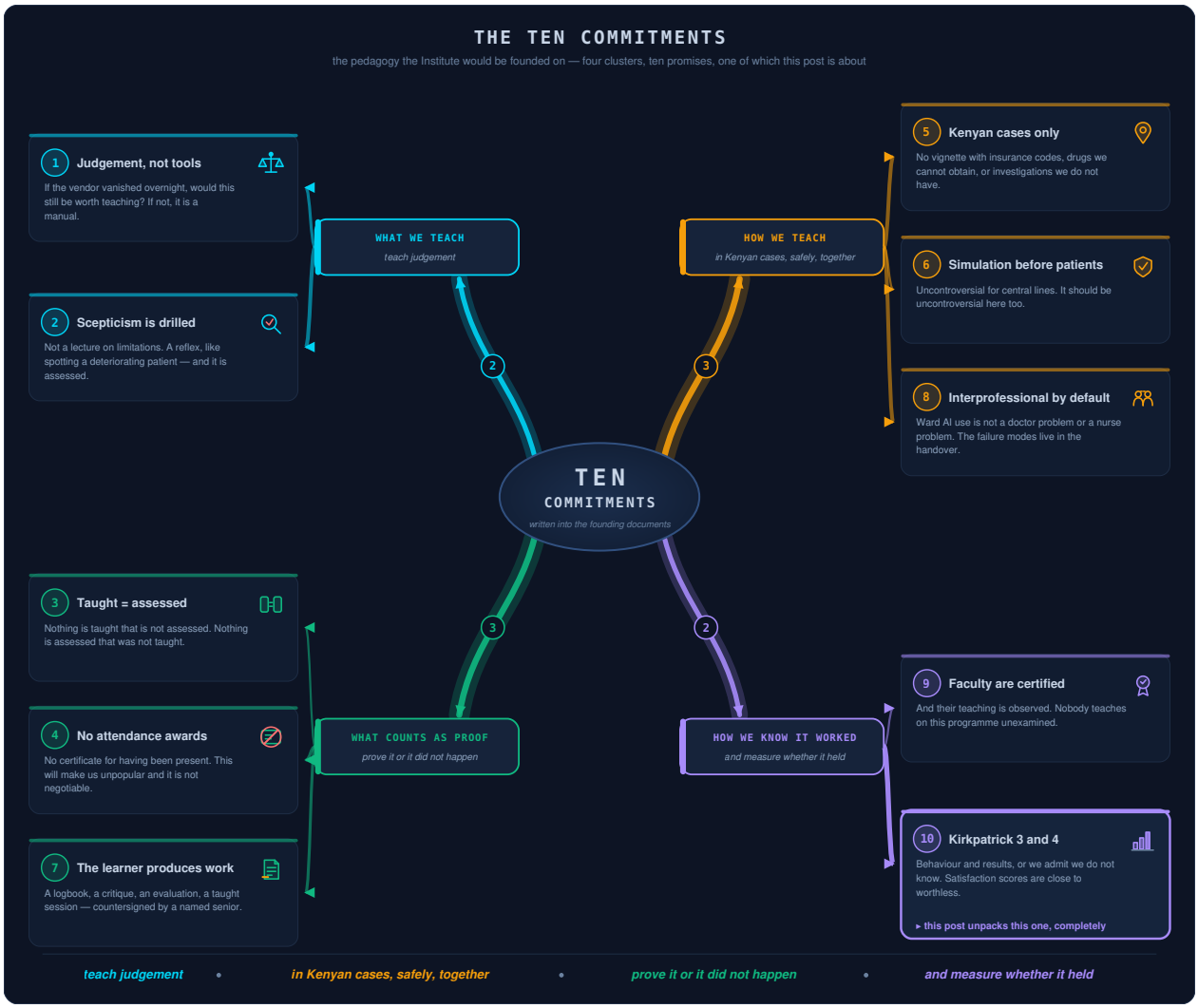
Your browser does not support the audio element.

Commitment ten, in the blueprint, is one sentence:

> **We measure at Kirkpatrick 3 and 4, or we admit we do not know.** Satisfaction scores are close to worthless.

It is the shortest of the ten and the one most likely to be quietly dropped, because it is the only one that costs money after the teaching has finished and everybody has gone home pleased with themselves. So it is worth setting out, at length and from first principles, exactly what it commits us to.

First, the whole set, because commitment ten does not stand alone — it is the last of four clusters, and the other nine are what it is measuring.



The ten pedagogical commitments. Ten loose items sit right at the edge of working memory; four coloured groups do not. Read the phrases along the bottom as one sentence — ***teach judgement · in Kenyan cases, safely, together · prove it or it did not happen · and measure whether it held****. The tenth is the subject of everything below.*

Part 1 — What Kirkpatrick actually is

1.1 The problem it was invented to solve

In 1954 a doctoral student at the University of Wisconsin called Donald Kirkpatrick wrote a dissertation on how you would tell whether a training course had done anything. In November 1959 he turned it into a four-part series of articles for the journal of what was then the American Society of Training Directors, one article per idea. He was writing about supervisors in American industry, not clinicians, and he was not trying to build a theory. He was trying to stop people claiming success on the basis of the form the delegates filled in at the end of the day. (Kirkpatrick Partners' own account; a useful historical corrective on the attribution is Will Thalheimer's.)

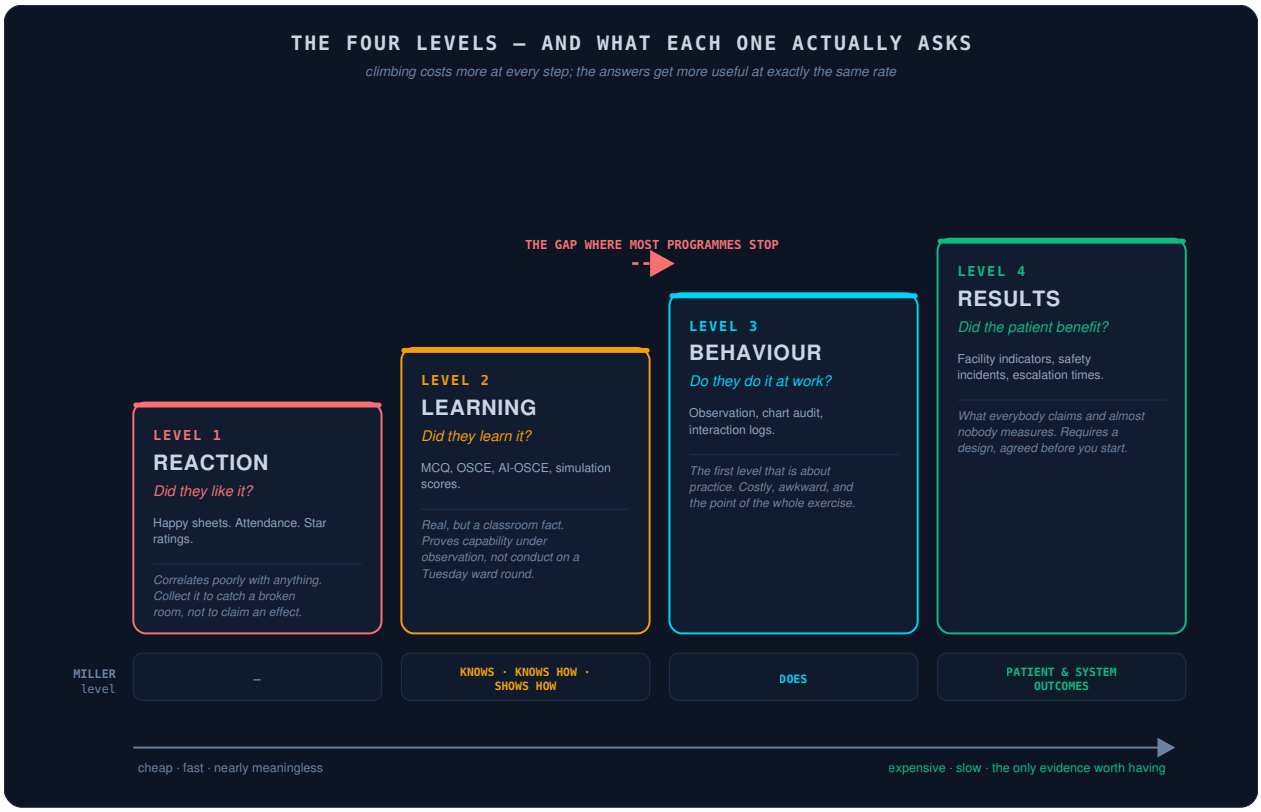
Seventy years later the four articles have hardened into the default vocabulary of training evaluation across every industry, including ours. That ubiquity is a problem in itself, because a framework everyone recites is a framework nobody examines.

1.2 The four levels, in plain language

Here is the whole model. If you have never planned a course in your life, this is all you need to start with.

Imagine you have just run a two-day workshop for twenty clinical officers. There are four completely different questions you could ask about it, and they are not four ways of asking the same thing. They are four different questions with four different answers, and it is entirely possible for the answer to be *yes* at one level and *no* at the next.

1. **Did they like it?** — *Reaction.* You find out by asking them.
2. **Did they learn it?** — *Learning.* You find out by testing them.
3. **Do they do it at work?** — *Behaviour.* You find out by going to their workplace, months later, and looking.
4. **Did anything change for the patient?** — *Results.* You find out by measuring something about the service, and by having designed the measurement before you started.



The four levels as a staircase. Each step costs more to climb than the one below it, in money, in time, and in the number of people whose cooperation you need. Underneath each is the corresponding level of Miller's pyramid — the standard way clinical educators describe what an assessment is actually evidence of.

> **The single most important thing to understand about the levels.** They are not a quality ranking of *evidence*. They are a ranking of *what question you are entitled to answer*. A beautifully conducted Level 2 study is excellent evidence about Level 2 and no evidence at all about Level 3. The commonest error in medical education — and I have made it myself — is to run a rigorous Level 2 assessment and then write a discussion section about practice change.

1.3 Why almost everybody stops at two

Because levels 1 and 2 happen in the room, while you still have everyone's attention and a budget line, and levels 3 and 4 happen somewhere else, months later, after the funder's report is due.

Level 1 costs a sheet of paper. Level 2 costs an exam and someone to mark it. Level 3 costs a trained observer travelling to a facility, an ethics approval, a data-sharing agreement, a clinician's afternoon, and a statistician. Level 4 costs all of that plus a study design agreed before the intervention starts, which means you have to have been thinking about evaluation at the moment you were most excited about the teaching.

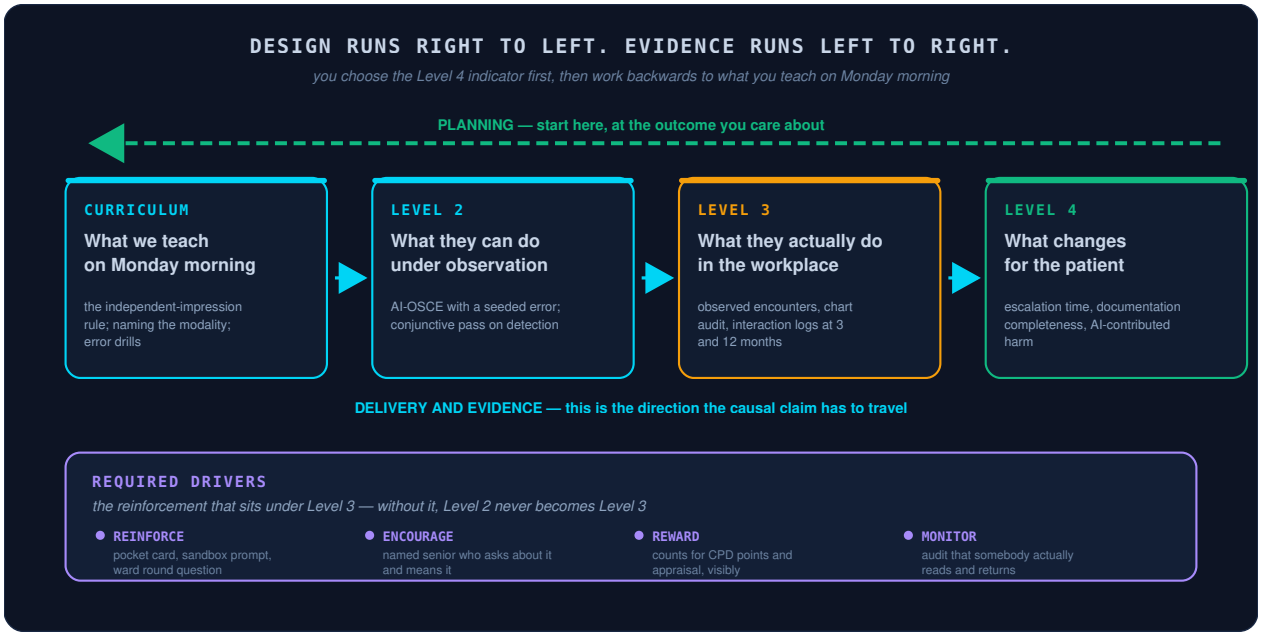
That asymmetry — not laziness, not dishonesty — is why the literature on educational interventions is a very tall pile of Level 2 studies. It is also why any institution that intends to do better has to write the commitment down *in advance*, in a founding document, where breaking it requires a public argument. Hence commitment ten.

1.4 The 2016 update: plan backwards, and build the drivers

In 2016 James and Wendy Kirkpatrick published what they call the **New World Kirkpatrick Model**, and it fixed the two things that were most often got wrong.

Plan backwards. Do not design a course and then wonder how to evaluate it. Choose the Level 4 result you are trying to move, define the Level 3 behaviours that would plausibly move it, define the Level 2 capabilities those behaviours require, and only then write the curriculum. Evaluation stops being an appendix and becomes the design brief.

Build the required drivers. This is the more useful of the two. The New World model names the thing that sits *between* Level 2 and Level 3 — the reinforcement, encouragement, reward and monitoring in the workplace without which a trained capability simply never becomes a habit. If you teach a clinician something on Friday and nobody on the ward ever mentions it again, you have not built a behaviour. You have built a memory, and memories decay.



Design runs right to left; evidence runs left to right. You choose the patient-level outcome first and derive the curriculum from it, but the causal claim has to travel the other way, through each stage, and each arrow is a place where it can fail. Underneath, the four required drivers — the reinforcement without which Level 2 never becomes Level 3.

> For our purposes the required drivers are not a footnote. My honest expectation is that the drivers will turn out to matter more than the curriculum. Two facilities receiving identical teaching, one of which has a consultant who asks "what did you think before you asked it?" on every ward round and one of which does not, will produce different Level 3 data — and the difference will have nothing to do with the course.

1.5 The criticisms, which are serious and which I accept

I would not want to build an institution on a 1959 industrial training model without saying plainly what is wrong with it.

It smuggles in a causal chain that may not exist. The levels are usually presented as a hierarchy in which each level causes the next. There is no strong empirical basis for that. Someone can change behaviour without having scored well on the test, and can score well without changing anything. Treating the hierarchy as a causal ladder is an assumption, not a finding.

It was built for simple interventions with short endpoints. Yardley and Dornan's 2012 critique in *Medical Education* is the one to read. Their conclusion is uncomfortable and, I think, correct: the levels carry so many implicit assumptions that they suit only relatively simple instructional designs, short-term endpoints, and beneficiaries other than the learner — conditions met by perhaps a fifth of medical education evidence reviews. Applied outside those conditions as a critical-appraisal tool, the hierarchy adds little and can actively mislead.

It treats the learner as a means. Level 4 is defined as benefit to the organisation. A clinician who becomes a better, more sceptical thinker has gained something the model has no box for.

So why use it? Because it is a shared vocabulary that a hospital board, a professional council, a funder and a clinician-educator can all read without a glossary — and because the specific failure it was invented to prevent (declaring victory on the strength of a happy sheet) is precisely the failure most likely to occur here. I use it as a **checklist against self-deception**, not as a theory of learning. Where it does not fit, I would say so in the evaluation report rather than force the data into it.

1.6 Two extensions worth knowing

Two adaptations are standard in health professions education and both improve on the original for our purposes.

Barr, Freeth and Hammick's split levels, developed for interprofessional education, subdivide two of the four: Level 2a (modification of attitudes and perceptions) and 2b (acquisition of knowledge and skills); Level 4a (change in organisational practice) and 4b (benefit to patients). Given commitment eight — interprofessional wherever the work is interprofessional — this split is the one I would actually adopt. Attitude towards AI and skill with AI are different things and they can move in opposite directions. So can organisational practice and patient benefit.

Moore's expanded outcomes framework stretches the four into seven for continuing medical education: participation, satisfaction, learning (declarative and procedural), competence, performance, patient health, community health. Its virtue is the explicit gap it opens between **competence** (level 5's neighbour — can do, in a controlled setting) and **performance** (does do, in practice). That gap is exactly where the AI-OSCE ends and workplace assessment begins, and CPD frameworks in several jurisdictions are built on it.

> **Vocabulary, once, so the rest of this post is unambiguous.**
> **Competence** — what a clinician can do when they know they are being assessed. Measured

> by examination and simulation. Kirkpatrick 2. Miller's *shows how*.
> **Performance** — what a clinician does when nobody has told them it counts. Measured in the workplace. Kirkpatrick 3. Miller's *does*.
> The distance between them is not a defect in the assessment. It is a real and permanent feature of professional practice, and the entire argument of this post is that you have to go and measure it rather than assume it away.

Part 2 — Applying this to student-clinicians: levels 1 and 2

From here on, everything is concrete. The cohort I have in mind is a group of student-clinicians and early-career clinicians — medical officers, clinical officers, nurses and midwives — completing Level 1 of the Institute's common core, the module that teaches the Clinical 4Ds: Delegation, Description, Discernment and Diligence.

2.1 Level 1 — what we collect, and the one thing it is good for

We collect reaction data. We collect it on the day, it takes four minutes, and we largely ignore it. But "largely" is not "entirely", and it is worth being precise about the exception, because the blanket dismissal of Level 1 is its own kind of sloppiness.

What reaction data cannot tell you: whether anything was learned, whether anything changed, or whether the teaching was any good. Learner satisfaction and learning outcome are weakly and sometimes inversely related. Effortful teaching that produces durable learning frequently feels worse in the room than fluent teaching that produces nothing — the well-documented gap between how well people think they are learning and how well they actually are.

What it can tell you: whether something was *broken*. A room where nobody could hear. A simulation that crashed. A facilitator who was hostile. Those are real, actionable, and invisible in the exam data. So the form we would use has almost nothing on it about enjoyment and four questions of the New World "relevance and commitment" type:

- What is one thing you will do differently on your next shift?
- What is one thing that will make that difficult?
- Was there anything in the two days you could not follow?
- Was there anything that did not work — room, kit, sandbox, materials?

Question one is a free-text field that turns into a Level 3 hypothesis. Question two is a free-text field that turns into a required-driver specification. Neither is a satisfaction score.

> **The rule I would enforce.** No Level 1 result ever appears in a report, a board paper, or a funding application as evidence of effect. Not once, not with a caveat. The moment "97% of participants rated the course highly" is permitted into an outcome section, commitment ten is dead, because that sentence is always available and always cheaper than the truth.

2.2 Level 2 — what "learning" means when the thing taught is a habit of mind

Level 2 is where most of the Institute's assessment machinery lives, and I have written about the two hardest parts of it elsewhere: what an AI-OSCE is and why one station hides a deliberate error, and how the pass mark is set by a modified Angoff panel rather than an arbitrary 50%. I will not repeat those here. What matters for the present argument is the shape of the claim a Level 2 pass licenses.

The Institute's Level 2 evidence has three components:

A 40-item invigilated knowledge test, weighted 40% towards Discernment, with a cut score set by panel. This establishes *knows* and, for the reasoning items, *knows how*.

An AI-OSCE, in which the candidate consults with a standardised patient while an AI system is available in the sandbox, and in which some stations seed a clinical error into the AI's output. Scored on delegation, description, error detection and correction, and documentation and disclosure. Error detection is a **conjunctive** requirement — you cannot pass by compensating elsewhere, in the same way a candidate cannot compensate for a fatal drug error with excellent communication. This establishes *shows how*.

A countersigned portfolio product — a logbook, a critique of a real AI-assisted decision, a taught session — read and signed by a named senior person, per commitment seven.

> **What a Level 2 pass entitles us to say:** this candidate, on a specified day, in a simulated encounter, in the knowledge that they were being assessed, detected a seeded error and documented it appropriately.
>
> **What it does not entitle us to say:** anything whatsoever about what they will do at 3 a.m. in a busy casualty department in eight months' time, when the model is fluent and confident and they are tired and the queue is thirty deep.

That second paragraph is the whole reason Level 3 exists, and it is not a hypothetical worry. The 2025 NEJM AI randomised trial that motivated the entire blueprint found physicians who had already completed twenty hours of AI-literacy training still deferring to deliberately erroneous model output. Those physicians would, I have no doubt, have passed a knowledge test on automation bias. The knowledge was not the binding constraint. That is a Level 2/Level 3 dissociation, observed directly, in exactly our population, on exactly our topic.

Part 3 — Level 3, in operational detail

Here is what the blueprint says, in full, and what the rest of this section unpacks:

> **Level 3 — behaviour.** At three and twelve months post-training: workplace-based assessment by a trained observer; chart audit for documentation of AI-assisted decisions; and, with consent and appropriate governance, sandbox interaction logs showing whether the independent-impression rule survived contact with real work. My working hypothesis — which I would want tested and would not be surprised to see refuted — is that the independent-impression discipline decays fastest and needs the earliest booster.

Four claims are packed into that paragraph: a set of target behaviours, a schedule, three data sources, and a falsifiable hypothesis. Take them in order.

3.1 First, name the behaviours — or you are not measuring anything

You cannot measure "behaviour". You can only measure *specified behaviours*, and specifying them is most of the work. A Level 3 plan that says "we will assess whether they apply their learning" is not a plan; it is a sentence that sounds like a plan.

Here are the four I would specify, in descending order of how much I care about them.

B1 — The independent-impression rule. *Before opening the model on a diagnostic question, the clinician forms and records their own working impression.* This is the load-bearing behaviour of the entire curriculum, so it is worth being exact about why.

Automation bias is not principally a failure of knowledge. It is an anchoring effect. Once a fluent, confident, well-formatted differential is on the screen, the clinician's own reasoning is no longer independent of it — it is a revision of it. Every subsequent thought is conducted in the model's frame. The systematic review by Goddard, Roudsari and Wyatt puts automation bias errors at roughly 6–11% of cases in decision-support studies, in both directions: errors of *commission* (following incorrect advice) and errors of *omission* (failing to act because the system did not prompt). Commission errors, they found, arise from a combination of not attending to available contradictory information and a belief in the superior judgement of the automated aid.

The rule is the countermeasure, and it works by sequence rather than by effort. It does not ask the clinician to be more sceptical, which is not a thing a person can reliably do on demand. It asks them to **commit to a position before the anchor arrives**, which converts an unfalsifiable intention into an observable act with a timestamp. That is precisely what makes it measurable — and it is why I would put it first.

B2 — Naming the modality. *Before delegating, the clinician can say which of automation, augmentation or agency they are operating in.* Almost every serious failure mode I

can construct involves someone operating in agency mode while believing they are in augmentation mode. Twenty minutes of teaching; disproportionate yield.

B3 — Documentation and disclosure. *Where AI materially contributed to a clinical decision, the record says so, says how, and says what the clinician did about it.* This is also the behaviour with the clearest legal and governance load — see The Law Is Part of the Architecture for why, under the Digital Health Act and its data-governance requirements, this is not merely good manners.

B4 — Escalation unchanged by model reassurance. *A clinician who would have escalated on clinical grounds still escalates when the model is reassuring.* The hardest to observe and, if it fails, the one that kills someone.

> Notice that every one of these is written as an observable act with an actor, a trigger and a verb. “Demonstrates appropriate scepticism” is not on the list, because two trained observers cannot reliably agree on whether it happened. If you cannot write the behaviour in a form where two observers would agree, you cannot assess it, and you should either rewrite it or drop it. That constraint disciplines the curriculum as much as the evaluation.

3.2 The schedule, and why three and twelve

Three months and twelve months are not arbitrary, and they are not simply convenient.

The retention literature in procedural and resuscitation skills is reasonably consistent on shape even where it disagrees on magnitude. A 2021 systematic review of retention after simulation training found significant decline in performance scores as early as three months, with scores nonetheless remaining above baseline at three and six months — decay, then partial plateau, rather than a return to zero. The advanced life support literature reports knowledge and skills decaying by six months to a year, with **skills decaying faster than knowledge**.

That last finding is the one I would generalise from, cautiously. If skills decay faster than knowledge in resuscitation, then in our setting the *procedural discipline* (do this before that) should decay faster than the *declarative content* (what a language model is). Three months is early enough to catch the first slope and still act on it. Twelve months tells you whether anything survived a year of real work, staff rotation, and a new model version. Six months would be better than nothing but is the least informative single point — it lands in the plateau, where the curve is flattest and least diagnostic.

*The hypothesis, drawn. These curves are **illustrative and not data** — they are what commitment ten predicts, laid out so that it can be checked. If the real curves come back in a different order, the hypothesis is wrong and the booster goes somewhere else. The dashed line is the acceptable-practice threshold, which has to be set before the first measurement rather than drawn around whatever we happen to find.*

3.3 Data source one — workplace-based assessment by a trained observer

What it is. A senior clinician sits in on a real consultation, watches, and completes a short structured form immediately afterwards, with feedback to the trainee. The two standard instruments are the **mini-CEX** (mini Clinical Evaluation Exercise, for whole encounters) and **DOPS** (Direct Observation of Procedural Skills, for procedures). Both are cheap, brief, and designed for repetition rather than perfection.

How it would be adapted here. A conventional mini-CEX form scores history, examination, professionalism, clinical judgement, organisation. Ours would add four domains mapped directly onto B1–B4, each with an anchored scale and each with a free-text justification field. The critical design decision is that **the observer records the sequence**, not just the quality: did the clinician form an impression before opening the model, or after? That is a binary with a timestamp, not a judgement call, and it is the single most valuable field on the form.

How many, and why it matters. This is where most workplace-based assessment schemes fall over, so the numbers deserve stating. A single mini-CEX is a very noisy measurement. Reliability accumulates across encounters and across assessors: work on the mini-CEX has reported reliability around 0.73 when aggregating roughly fifteen encounters, and the composite reliability literature shows that a *portfolio* combining instruments gets there faster — a coefficient near 0.80 from a combination of mini-CEXs, DOPS and multi-source feedback rounds, with fewer of each than any single instrument

would need alone. The practical implication for us is emphatic: **do not attempt to make a high-stakes judgement from one observation.** Aggregate.

What contaminates it. Three things, all documented, all worth designing against:

- **Hawthorne effect.** They behave well because you are watching. This is not a reason to abandon observation; it is a reason to read observation as a measure of *best-case* behaviour. If the independent-impression rule is not followed even when a consultant is sitting in the corner with a clipboard, you have learned something extremely important.
- **Assessor stringency.** In published analyses of mini-CEX score variance, examiner stringency has been found to account for a substantial share — around 29% in one dataset, against roughly 13% for the trainee's own aptitude for the attachment. Which is to say: *who assesses you can matter more than how good you are*. The mitigations are assessor training, multiple assessors per trainee, and never using a single assessor's score as a gate.
- **Case mix.** A straightforward case gives the trainee no reason to consult a model at all, and therefore no opportunity to demonstrate B1. The sampling frame has to specify diagnostic-uncertainty encounters, or the instrument measures nothing.

> **Operational note.** I would specify: minimum eight observed encounters across the twelve months, at least four different assessors, at least six encounters flagged in advance as involving diagnostic uncertainty, all assessors having completed the faculty certification of commitment nine and having had at least one of their own observations observed. Assessor-level score distributions published internally each quarter, so that a systematically lenient or harsh assessor is visible without anyone having to make an accusation.

3.4 Data source two — chart audit

What it is. Structured retrospective review of the clinical record against explicit criteria. It is unobtrusive, it scales, it covers everyone rather than a sample of the willing, and it is the workhorse of quality measurement in health systems everywhere. It has been used to evaluate CME programmes precisely because it reaches Level 3 without requiring anyone to be observed — see, for example, this study using chart review to evaluate a CME programme.

What we would audit. For each sampled encounter: is there a recorded working impression, and is it timestamped before the AI interaction? Is AI involvement documented at all? Is the nature of the contribution described? Where the model's output was not followed, is the reasoning recorded? Where a decision was escalated or not escalated, is the rationale there?

How to do it without fooling yourself. Chart audit is easy to do and easy to do badly. The design points that matter:

- **Specify the sampling frame before you look.** Consecutive encounters in defined windows, stratified by clinician and by presentation type. Not “cases the clinician chose to submit”, which measures self-presentation.
- **Blind the auditors** to whether the clinician has completed training, and to the audit period, as far as the record permits.
- **Double-code a fraction** — 15–20% is conventional — and report inter-rater agreement with a chance-corrected statistic (Cohen’s or Fleiss’s kappa). An audit without a reported kappa is an opinion with a denominator.
- **Pilot the codebook** on twenty records and expect to rewrite half of it. Every ambiguity you find in the pilot is an ambiguity that would otherwise have become noise.

> **The limitation that cannot be designed away.** Chart audit measures documentation, not thought. A clinician who has completely surrendered their judgement to the model can write an impeccable note recording an independent impression they did not actually form. If documentation improves and observed behaviour does not, the honest interpretation is that we have taught documentation — and that is a real finding, and a curriculum problem, and it must not be reported as behaviour change.

3.5 Data source three — sandbox interaction logs

What it is. The Institute’s teaching platform is model-agnostic by architecture, and its sandbox can record the sequence of interactions: what was asked, when, in what order, and what happened next. That makes it the only one of the three sources that directly observes **order of operations** — which, for B1, is the entire measurement.

What it can establish. Whether a working impression was entered before the first model query. Time-to-first-query from the start of the encounter. Whether the clinician queried again after receiving an answer, or accepted it. Whether outputs flagged as uncertain were treated differently from confident ones. These are behavioural traces of exactly the discipline we are trying to instil, collected without an observer in the room and therefore without a Hawthorne effect.

What it must never become. This is the point in the design where an evaluation turns into surveillance if nobody is paying attention, and I would want the constraints written into the founding instruments alongside the independence rules:

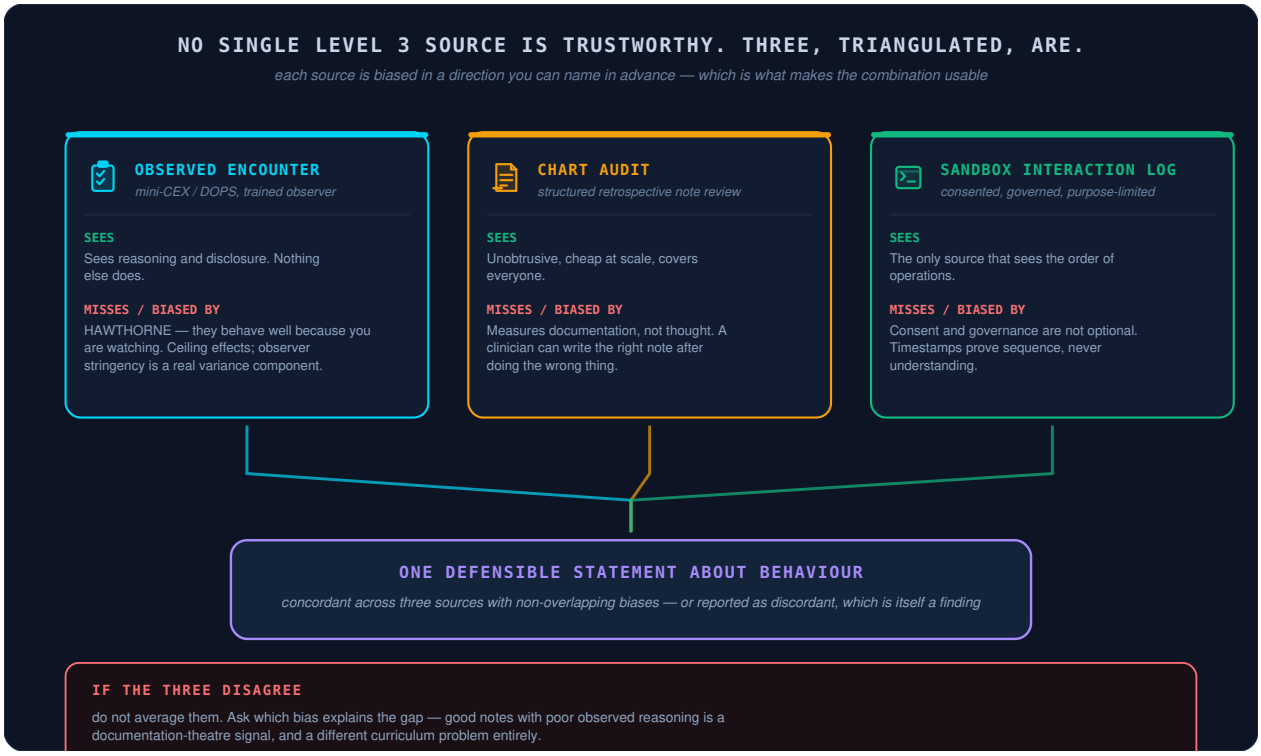
- **Explicit, specific, revocable consent** to log analysis for evaluation, separate from consent to use the platform, and refusable without any effect on certification.
- **Purpose limitation in writing.** Logs are analysed for aggregate evaluation. They are not used for individual performance management, not shared with employers, and not admissible in a disciplinary process. If that undertaking cannot be given and kept, the logs should not be collected.

- **Minimisation and pseudonymisation** at the point of collection, with a retention period and a deletion date, consistent with the Digital Health Act's data-governance provisions and the Data Protection Act.
- **Governance approval and publication** of the analysis protocol before any analysis is run.

> A clinician who suspects the logs are being read by their employer will change their behaviour — and the behaviour they will adopt is *performing the ritual*: typing an impression they have not formed, because the system is watching. At that point the instrument has not merely stopped measuring the behaviour, it has actively destroyed it. Trust is not an ethical nicety here; it is a measurement precondition.

3.6 Triangulation — why three sources and not one

None of the three is trustworthy alone. Each is biased in a direction you can name in advance, and — this is the point — the directions do not coincide.



Three sources with non-overlapping biases. The observed encounter sees reasoning but is contaminated by being observed; the chart audit is unobtrusive but sees only what was written; the log sees sequence but never understanding. Agreement across all three is worth far more than a strong result from any one — and disagreement is itself informative, provided you resist the urge to average it away.

The discipline this imposes is worth stating explicitly, because it is where evaluations usually go soft: **decide, before you collect anything, what you will conclude from each pattern of agreement and disagreement.** Write it down. Four cases:

Observation	Chart	Log	Reasonable reading
Good	Good	Good	The behaviour is established. Report it, publish the effect size, and check again at twelve months.
Poor	Good	Poor	Documentation theatre. We have taught note-writing, not reasoning. Curriculum problem, and a serious one.
Good	Poor	Good	The behaviour exists; the record does not reflect it. A documentation and workflow problem — real, but a different fix.
Good	Good	Poor	Look hard at the log analysis before you believe it. Sandbox use may simply not reflect real workflow if clinicians have moved to a consumer tool on their own phone — which is itself the most important finding in the study.

That last row deserves emphasis. If trained clinicians abandon the governed sandbox for an ungoverned consumer chatbot, every instrument above is measuring the wrong system, and the correct response is not a better log analysis but an urgent conversation about why the sanctioned tool lost.

3.7 The hypothesis, and what would refute it

The blueprint states a working hypothesis: **the independent-impression discipline decays fastest and needs the earliest booster.** I want to be precise about its status, because a hypothesis you cannot lose is not a hypothesis.

The reasoning behind it. B1 is procedural rather than declarative; procedural skills decay faster in the retention literature. It is also the behaviour with the highest immediate cost to the clinician — it takes thirty seconds *before* the shortcut, at the exact moment the shortcut is most tempting, and the pressure to skip it rises with queue length. And it is the least visible to colleagues: nobody can tell from the outside whether you formed an impression first, so social reinforcement is weak. Three independent reasons to expect fast decay.

What would refute it. If, at three months, the observed and logged rate of independent impression before first query is not significantly lower than at course exit — or is not lower than the corresponding rates for B2 and B3 — the hypothesis is wrong. It is also refuted, differently, if all four behaviours decay at indistinguishable rates, which would mean the specificity of the claim was unfounded and that boosters should be general rather than targeted.

What we would do if refuted. Move the booster. The point of a stated hypothesis is that being wrong is cheap and informative, provided you said it out loud first. Which is why it is in the blueprint rather than in a drawer.

> **On the booster itself.** The correct response to decay is not to repeat the course. Repetition of the original material is the least efficient intervention available. The evidence on retention favours brief, spaced, effortful retrieval over massed re-teaching — the testing effect, which is one of the most robust findings in the learning sciences. So: 45 minutes, one seeded-error case, a required retrieval attempt before any teaching,

> feedback, done. Timed to land just after the three-month measurement, so that the measurement is uncontaminated and the intervention is still early enough to matter.

Part 4 — Level 4, in operational detail

Again, the blueprint text in full:

> **Level 4 — results.** Facility-level indicators agreed in advance: documentation completeness, appropriate investigation rates, time-to-escalation for deteriorating patients, and incidents in which AI contributed to harm. Where we can run a stepped-wedge design across facilities, we should. Where we cannot, we should report the limitation honestly rather than implying causation from a before-and-after chart.

Three things are being committed to: indicators fixed in advance, a randomised design where feasible, and honesty about causation where it is not.

4.1 "Agreed in advance" is the load-bearing phrase

Everything else in that paragraph is technique. This is the part that determines whether the evaluation is worth anything.

If indicators are chosen after the data are in, you will choose the ones that moved. Not through dishonesty — through the ordinary human process of finding the favourable comparison more interesting than the unfavourable one, and of constructing a plausible story about why it was the right measure all along. The published literature on selective outcome reporting is unambiguous that this happens routinely in fields staffed by careful, well-intentioned people.

So: indicators, definitions, numerators, denominators, analysis method, subgroups and stopping rules, all written down and registered before the first facility is trained. Preferably published. If the pre-registered analysis produces a null result, the null result is the finding.

4.2 The four indicators, operationalised

Naming an indicator is not defining it. Each of the four needs a numerator, a denominator, a data source and an anticipated failure mode.

Documentation completeness. *Numerator:* encounters in which AI involvement is documented with nature of contribution and clinician action. *Denominator:* encounters in which the interaction log shows an AI interaction occurred. *Source:* chart audit linked to sandbox log. *Failure mode:* the easiest indicator to move by exhortation alone, and therefore the weakest evidence of anything that matters. Treat a large improvement here with suspicion, not celebration.

Appropriate investigation rates. *Numerator:* investigations ordered that meet prespecified appropriateness criteria for the presentation. *Denominator:* all investigations ordered for that presentation. *Source:* chart audit against a criteria set agreed by a clin-

ical panel before the study. *Failure mode:* "appropriate" is a judgement, and the panel that defines it is doing standard-setting — the same problem, with the same solution, as setting a cut score. Note also that this indicator is **directionally ambiguous**: AI can drive both over-investigation (defensive prompting on a long differential) and under-investigation (false reassurance). The pre-registration must state which direction constitutes improvement for which presentation, or the indicator is unfalsifiable.

Time-to-escalation for deteriorating patients. *Numerator/measure:* median minutes from first recorded abnormal early-warning score to documented senior review. *Denominator:* patients meeting the deterioration trigger. *Source:* observation charts and clinical record. *Failure mode:* highly sensitive to staffing, bed state and time of day. Requires adjustment and adequate volume; in a small facility, a handful of night shifts can swing it.

Incidents in which AI contributed to harm. *Numerator:* reported incidents where structured review judges AI to have contributed. *Denominator:* admissions or encounters. *Source:* incident reporting plus mortality and morbidity review. *Failure mode:* rare-event counting with a reporting rate that the intervention itself will change. Training people to notice AI-related harm will increase reported AI-related harm. **A rise in this indicator after training may be a success, not a failure**, and the pre-registration has to say so in advance, or the first honest facility will be punished for its honesty.

> **Balancing measures.** Any serious Level 4 set includes indicators that would detect the intervention doing harm. Ours: consultation duration (are we making every encounter slower?), clinician-reported cognitive load, referral rates (are we shifting risk upwards rather than managing it?), and — the one I would most want to see — whether sanctioned-tool use falls while overall AI use does not, which would mean we had driven the behaviour underground.

4.3 The stepped-wedge design

The attribution problem at Level 4 is severe. Facility indicators move for a hundred reasons — a new clinical officer, a drug stock-out, a change in referral patterns, a national guideline, a rainy season. A before-and-after comparison cannot distinguish any of that from your training.

A **stepped-wedge cluster randomised trial** is the design that fits this situation almost too well.

Five facilities, six periods. Every facility begins as a control and every facility ends up trained; what is randomised is the order, not the receipt. Each facility acts as its own control before crossover, and contributes to the concurrent control group for the facilities that have not yet crossed. The staircase is the wedge.

Why it fits. You are going to roll the programme out to every facility anyway — the Institute's whole purpose is national coverage. A parallel-arm trial would require withholding training from half the facilities indefinitely, which is neither politically nor ethically viable. The stepped wedge randomises only the sequence. Nobody is denied anything; they are asked to wait a defined and randomly allocated number of months. That is a very different conversation with a hospital superintendent, and it is the reason the design has become common in health systems research.

What it costs you, honestly. Time and treatment effect are confounded by construction, because later periods contain more trained facilities. The analysis must include a fixed effect for period, and it depends on the assumption that secular trends are common across clusters. It is also, as the CONSORT extension for stepped-wedge trials sets out, potentially at *greater* risk of certain biases than a parallel cluster trial — within-cluster contamination in particular, since every cluster experiences both conditions. The extension exists precisely because these trials were being reported without the information needed to judge them; the requirement to give a clear justification for choosing the design is the part I would hold us to hardest.

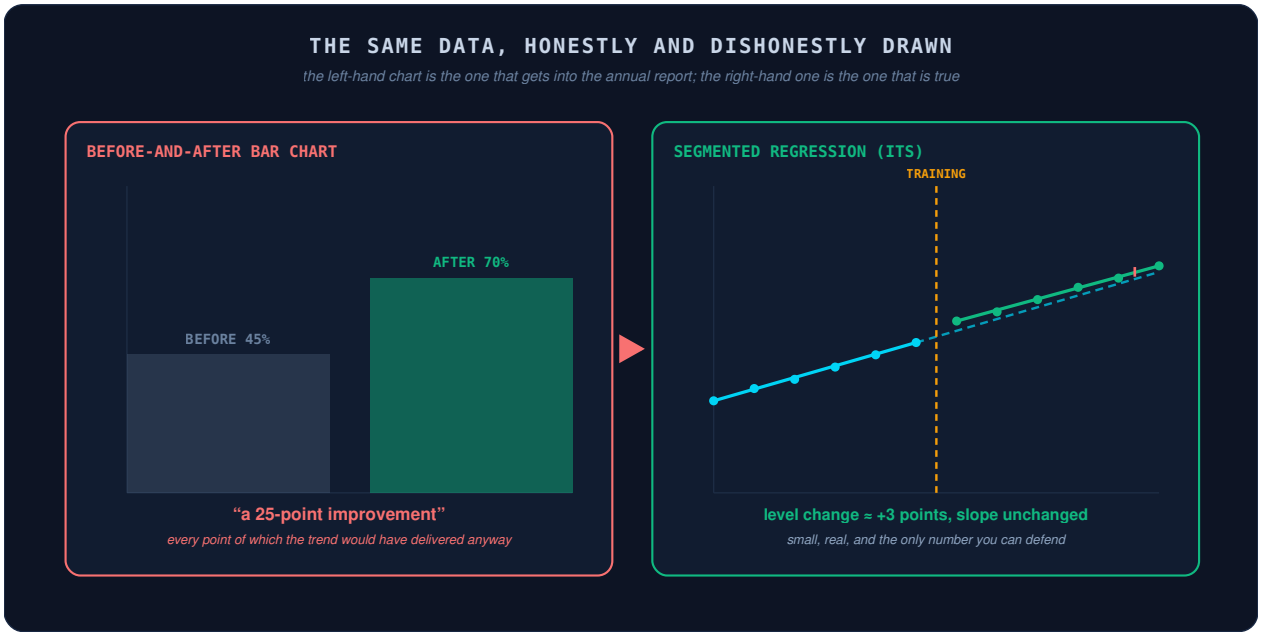
The practical parameters. You need enough clusters — below about four, the randomisation buys you very little and the analysis is fragile. You need an estimate of the intra-cluster correlation coefficient to power the study at all, and you almost never have a good one in advance, so you plan for a range and say what you assumed. You need to specify the transition period during which a facility is training and neither cleanly control nor cleanly intervention, and either exclude it or model it. And you need to think

hard about contamination: clinicians rotate between facilities in exactly the health systems where this design is attractive, and a rotating registrar carries the intervention across a cluster boundary in their head.

4.4 When you cannot run a wedge

Often you will not be able to. The rollout order may be decided by a ministry, by a funder, or by which facility has working connectivity. That is not a reason to abandon Level 4 — it is a reason to be explicit about what a weaker design can and cannot support.

The fallback is an **interrupted time series** analysed by segmented regression, with a concurrent control series where one exists. The Cochrane EPOC standard is at least three data points before and three after; more is much better, and monthly points over two years either side is a reasonable target. The method estimates two things a before-and-after chart cannot: a **level change** at the intervention point and a **slope change** afterwards.



Identical underlying data, drawn twice. On the left, the chart that gets into the annual report: two bars, a 25-point improvement, and no way to see that the indicator was already climbing steadily before anyone was trained. On the right, the same series with its pre-existing trend made visible and extrapolated — leaving a real but modest level change of about three points, which is the only number anybody can defend.

This figure is, to me, the single most useful thing in the post. The left-hand chart is not a fabrication; every number in it is true. It is the *shape of the presentation* that manufactures the claim. Adding a control series strengthens it further, moving the analysis towards a difference-in-differences estimator that removes shocks common to all facilities — a national guideline change, a strike, a supply interruption.

And where even that is not available: say so. "We observed an improvement in documentation completeness from 45% to 70% over the period. We cannot attribute this to the training programme, because the indicator was already improving and we had no

control series.” That sentence costs nothing except the pleasure of a stronger claim, and it is the difference between an evaluation and an advertisement.

> It is entirely possible that a rigorous evaluation will show that some of what we teach does not change behaviour, or changes it in ways that do not benefit patients. If that happens, the correct response is to publish it and change the curriculum. **An institution that cannot survive its own negative findings is not an institution worth building.**

Part 5 — The other pedagogical instruments, and how each would be used

Kirkpatrick is a frame for asking questions. It contains no instruments. Everything that actually generates the evidence comes from somewhere else, and it is worth naming each tool, saying what it buys, and saying how it would be used here.

5.1 Miller's pyramid, and entrustment above it

What it is. George Miller's 1990 framework describes four levels of what an assessment is evidence of: **knows** (facts), **knows how** (applying them), **shows how** (demonstrating in a controlled setting), **does** (performing unobserved in practice). It is the most useful single diagram in assessment because it stops people confusing the bottom of the pyramid with the top.

How we would use it. As a blueprinting tool. Every assessment item in the Institute's programme is tagged with its Miller level, and the tags are published in the assessment blueprint. The rule that follows is commitment three's operational form: **no certificate is issued on *knows* alone**, and every certificate requires evidence at *shows how* and at least one countersigned data point at *does*.

The extension. Ten Cate and colleagues have proposed a fifth level — entrustment, or "trusted with future care" — which reframes the question from *how good is this trainee's performance* to *what would I now let this trainee do unsupervised*. That is the question a supervisor actually asks, and phrasing it that way tends to produce better-calibrated judgements than a numerical rating scale. For us the natural **entrustable professional activity** is: *"independently conducts an AI-assisted diagnostic consultation, including detection and disclosure of model error."* Rated on a supervision scale — observed only / with direct supervision / with indirect supervision / independently / able to supervise others — rather than on a 1-to-9 performance scale that no two assessors interpret alike.

5.2 Programmatic assessment

What it is. Van der Vleuten and Schuwirth's principle: no single assessment is ever adequate for a high-stakes decision. Instead, collect many low-stakes data points, each optimised for feedback rather than judgement, triangulate across methods, and have a committee synthesise them into the high-stakes decision when enough information has accumulated. Individual data points are maximised for learning; the decision is made on aggregate.

How we would use it. This is the organising principle for the entire Level 2/Level 3 apparatus, and it resolves the reliability problem from §3.3 elegantly. No single mini-CEX gates anything. No single chart audit gates anything. A **competence committee** — not the trainee's own supervisor — reviews the accumulated portfolio and makes a documented, reasoned progression decision, with the reasoning written down. That last part

matters more than it sounds: narrative quality in the record is what makes the decision defensible when it is challenged.

The failure mode to design against. Programmatic assessment collapses if the "low-stakes" data points are perceived as high-stakes. The moment trainees believe every mini-CEX is a judgement, they stop volunteering difficult cases and start volunteering easy ones, and the whole system measures case selection.

5.3 The AI-OSCE with a seeded error

Covered in full in One Hidden Error. Its role in the measurement architecture is specific and worth restating: it is the **only instrument that can create a known ground truth**. In the workplace you never know whether the model was right, so you cannot score detection. In a station where you planted the error, you know exactly what should have been caught, and the conjunctive requirement on the error-detection domain means it cannot be compensated away.

Its limitation is equally specific: it measures *shows how*, in a candidate who knows they are being assessed, and it is therefore an upper bound on real-world performance. Which is the entire argument for Part 3.

5.4 Standard setting

Every judgement above — passed, competent, acceptable, improved — requires a line, and a line requires a defensible process. The modified Angoff panel is how the knowledge test's cut score is set; borderline regression is the appropriate method for the OSCE, since the station data give you the borderline group directly.

The point I want to carry across into Level 4 is that this problem does not disappear when you move from exams to indicators. "Appropriate investigation rate" needs a standard as much as a 40-item MCQ does, and the same argument applies: a cut score is a policy decision about acceptable risk, and the only thing that makes it defensible is the quality and transparency of the process that produced it.

5.5 Retrospective pre-post, for the self-report you cannot avoid

The problem. Some things — confidence, perceived competence, self-reported frequency — can only be measured by asking. And a conventional pre-then-post self-report is systematically broken by **response-shift bias**: the course changes the learner's internal yardstick. A clinician who rated their AI competence 4/5 before the course may rate it 3/5 afterwards, having learned enough to know what they did not know. On a naive analysis, the course made them worse.

The fix. Ask both questions at the end. "Rate your competence now" and "thinking back, rate your competence before the course" — the *post-then-pre* design. Both ratings then use the same, post-course yardstick. The evidence is that this detects treatment effects that traditional pre-post analyses miss, though it introduces its own memory and social-desirability distortions and should never be the only measure of anything.

How we would use it. For confidence and self-efficacy only, always alongside an objective measure, and reported separately. The most interesting result would be a **divergence**: confidence rising while observed error detection falls is the signature of exactly the failure the whole programme exists to prevent, and it is a signal you can only get if you measure both.

5.6 A logic model

What it is. A one-page diagram of inputs → activities → outputs → short-term outcomes → long-term outcomes, with the assumptions on each arrow made explicit. Unglamorous, and the single highest-yield hour in the design of any evaluation.

How we would use it. As the artefact the Level 4 pre-registration is derived from. Its value is that it forces you to write the assumptions on the arrows. The arrow from "clinicians trained in the independent-impression rule" to "reduced time-to-escalation" carries at least four assumptions — that the rule survives to the workplace, that it changes what the clinician concludes, that the conclusion changes what they do, and that what they do is what determines escalation time. Writing them out tells you which are testable, which are heroic, and where the evaluation should look first when the result is null.

5.7 Audit and feedback, as a required driver

What it is. Measuring practice and giving clinicians the result. It is the most-studied behaviour-change intervention in health services research. The Cochrane review finds small but potentially important improvements in professional practice — and, more usefully, tells you *when* it works: when baseline performance is poor, when feedback comes from a supervisor or a respected colleague, when it is delivered more than once, when it is given both verbally and in writing, and when it includes explicit targets and an action plan.

How we would use it. This is where the evaluation stops being extractive. The chart audit of §3.4 is collected for evaluation; returning it to the clinician, with a target and an action plan, converts it into one of the required drivers. Same data, two functions. And a Level 3 measurement programme that gives nothing back to the people being measured will not survive contact with a busy clinical service, regardless of how good the design is.

5.8 Spaced retrieval, for the booster

The booster in §3.7 is a teaching intervention, and the retention literature is fairly clear about its shape. Effortful retrieval beats re-presentation; spacing beats massing; a test is a better learning event than a lecture. So the booster starts with a case and a required attempt *before* any teaching, rather than with a recap of the original slides. This is also why the interval matters: the booster is scheduled to land where the decay curve is steepest, which is precisely the information §3.2 is designed to give us.

5.9 The catalogue, in one table

Instrument	Kirkpatrick level	What it buys	How we would use it	Principal failure mode
Reaction form (relevance/ commitment type)	1	Detects broken delivery; generates L3 hypotheses	Four questions, none about enjoyment	Being cited as evidence of effect
MCQ with Angoff cut score	2	Defensible knowledge threshold	40 items, 40% weight on Discernment	Tests recall of scepticism, not scepticism
AI-OSCE with seeded error	2	Known ground truth; conjunctive error domain	Gate to certification	Upper bound only; candidate knows they are watched
Countersigned portfolio product	2→3	Forces production, not attendance	Named senior signs it	Signature becomes a formality
mini-CEX / DOPS	3	Sees reasoning and disclosure	≥8 encounters, ≥4 assessors, aggregated	Hawthorne; assessor stringency
Chart audit	3 / 4a	Unobtrusive, scales, covers everyone	Blinded, double-coded, kappa reported	Measures documentation, not thought
Sandbox interaction log	3	Order of operations, no observer effect	Consented, purpose-limited, aggregate only	Becomes surveillance; destroys the behaviour
Entrustment / EPA scale	3	Asks the question supervisors actually ask	Supervision-level anchors	Drifts to a performance rating in practice
Multi-source feedback	3	Interprofessional view of handover behaviour	Nurse and pharmacy raters included	Popularity contest without anchored items
Retrospective pre-post	2a	Corrects response-shift bias	Confidence only, alongside objective data	Memory and social desirability
Stepped-wedge CRT	4	Randomised counterfactual without denial	Where rollout order is ours to set	Confounding with time; contamination
Interrupted time series	4	Separates level change from pre-existing trend	Fallback; ≥3 points either side	Needs many points; no control series
Logic model	all	Makes assumptions on the arrows explicit	Derives the pre-registration	Written once, never revisited
Audit and feedback	driver	Turns measurement into reinforcement	Return the audit with a target and plan	One-off feedback with no action plan

Part 6 — What this apparatus still cannot tell you

Four honest limits, which I would want in the evaluation report rather than discovered by a critic.

It cannot establish that the training caused the patient outcome. Even a well-run stepped wedge across five facilities gives you an association under assumptions, with a confidence interval that will be wide. The causal chain from a two-day course to a mortality figure has at least six links and every one leaks.

It cannot measure the counterfactual clinician. We can measure what trained clinicians do. We cannot easily observe what the same clinician would have done untrained on the same patient. This is a limitation of the world, not of the design.

It will be confounded by the model changing under us. The system a cohort trained on in March is not the system they use in December. Model updates are a time-varying confounder that no educational design controls, and in a stepped wedge they are partially confounded with period. The honest response is to record model versions as a covariate and say plainly that it is a limitation.

It cannot capture what I would most like to know — whether a clinician has become a better thinker, or has merely acquired a compliant new ritual. B1 measured by timestamp is a proxy for a habit of mind, and a proxy is what it will remain. Anyone claiming otherwise is overselling.

What I would write into the founding documents

Compressed, so it fits on one page of an operations manual.

1. **No Level 1 result is ever reported as an outcome.** Reaction data informs delivery and generates hypotheses. Nothing else.
2. **Every certificate requires evidence at *shows how* and at least one counter-signed data point at *does*.** Nothing is issued on *knows* alone.
3. **Level 3 is measured at three and twelve months** on four pre-specified behaviours, from three sources with non-overlapping biases, aggregated across at least eight encounters and four assessors, with the analysis of concordance and discordance specified in advance.
4. **Interaction logs are consented, purpose-limited, pseudonymised, time-limited, and inadmissible in any individual performance process.** No exceptions, and the undertaking is published.
5. **Level 4 indicators, definitions, analysis and stopping rules are registered before the first facility is trained.** Balancing measures included. A rise in reported AI-related harm is pre-declared as potentially favourable.

6. **A stepped-wedge design is used wherever the rollout order is ours to set.** Where it is not, segmented regression with a control series where available, and an explicit statement that causation is not established.
7. **The evaluation protocol and its results are published regardless of outcome**, per the independence rules, and the external examiner sees the analysis before the board does.
8. **Every audit returns to the clinician** with a target and an action plan, within four weeks. Measurement that gives nothing back does not survive.

Coda

The reason commitment ten is written the way it is — *"or we admit we do not know"* — is that the second clause is the one that will actually be needed. Most of the time, for most of what we teach, a rigorous evaluation will return a wide confidence interval around a small effect, and the honest sentence will be that we cannot yet say.

That is not a failure of the evaluation. It is what evaluation is for. The alternative — a satisfaction score, a bar chart, and a claim — is available at all times, costs nothing, and tells you nothing about whether a clinician in a district hospital at three in the morning still forms their own impression before the machine offers one.

That is the only question worth answering. It is expensive to answer. Commitment ten is the promise to pay.

If you want the rest of the design: the full blueprint sets out the institution, the five tracks and the five gated levels; Borrowed From an Art School traces where the competency framework came from and what its licence permits; One Hidden Error covers the OSCE and the AI-OSCE in detail; The Angoff Panel covers standard setting; The Law Is Part of the Architecture covers the Kenyan legal and data-governance frame that Part 3.5 depends on; and AI Walks Into the Clinic is on how fast the ground is moving underneath all of it. Other writing is in the archive, and things I have built are under demos and lessons.

References

The model itself

- Kirkpatrick, D. L. (1959–60). Four-part series, *Journal of the American Society of Training Directors*. Summarised, with the later revision, in Kirkpatrick Partners, *An Introduction to the New World Kirkpatrick Model*.
- Thalheimer, W. *Donald Kirkpatrick was NOT the originator of the four-level model* — a useful corrective on attribution.
- Yardley, S. and Dornan, T. (2012). *Kirkpatrick's levels and education 'evidence'*. *Medical Education* 46(1):97–106. The critique to read before adopting the model.
- Moore, D. E. et al., expanded outcomes framework, levels 1–7 — see *A Conceptual Framework for Continuing Medical Education and Population Health*.
- Barr, H., Freeth, D., Hammick, M. et al., split levels 2a/2b and 4a/4b for interprofessional education — see the National Academies' *Measuring the Impact of Interprofessional Education*.

Assessment

- Miller, G. E. (1990). *The assessment of clinical skills/competence/performance*. *Academic Medicine* 65(9):S63–7.
- Ten Cate, O. et al. (2021). *Entrustment Decision Making: Extending Miller's Pyramid*. *Academic Medicine* 96(2):199–204.
- Schuwirth, L. and van der Vleuten, C. et al. (2019). *Programmatic assessment: can we provide evidence for saturation of information? Medical Teacher* 41(6):676–82. The originating idea is in van der Vleuten et al., *A model for programmatic assessment fit for purpose*, *Medical Teacher* (2012).
- Moonen-van Loon, J. et al. (2013). *Composite reliability of a workplace-based assessment toolbox for postgraduate medical education*. *Advances in Health Sciences Education*.
- Howard, G. S. et al. on response-shift bias; and *Controlling response shift bias: the retrospective pre-test design*, *Assessment & Evaluation in Higher Education* (2008).

Retention and decay

- Legoux, C. et al. (2021). *Retention of Critical Procedural Skills After Simulation Training: A Systematic Review*. *AEM Education and Training*.
- Yang, C.-W. et al. (2012). *A systematic review of retention of adult advanced life support knowledge and skills in healthcare providers*. *Resuscitation*.

Evaluation design

- Hemming, K., Taljaard, M., McKenzie, J. E. et al. (2018). *Reporting of stepped wedge cluster randomised trials: extension of the CONSORT 2010 statement with explanation and elaboration*. *BMJ* 363:k1614. Introduced in *Trials*.

- Ivers, N. et al. *Audit and feedback: effects on professional practice*, Cochrane Database of Systematic Reviews (2025 update of the 2012 review).
- Penfold, R. B. and Zhang, F. (2013). *Use of interrupted time series analysis in evaluating health care quality improvements*. *Academic Pediatrics* 13(6):S38–44.
- *Using chart reviews to evaluate a Continuing Medical Education (CME) program*, *BMC Medical Education* (2023) — a worked example of reaching Moore level 5 (performance) by retrospective record review.

Clinical AI and automation bias

- *Automation Bias in Large Language Model–Assisted Diagnostic Reasoning among Physicians Trained in AI Literacy — A Randomized Clinical Trial*, *NEJM AI*, 2025. The trial the blueprint is built around.
- Goddard, K., Roudsari, A. and Wyatt, J. C. (2012). *Automation bias: a systematic review of frequency, effect mediators, and mitigators*. *JAMIA* 19(1):121–7.
- OpenAI and Penda Health, *Pioneering an AI clinical copilot*, the underlying real-world study, and the critical reading in STAT News.
- AAMC, *Artificial Intelligence Competencies Across the Learning Continuum*; WHO, *Ethics and governance of AI for health: guidance on large multi-modal models*.

In the spirit of the framework's own Diligence competency: this post was drafted with AI assistance. The argument, the design decisions, the operational specifications and the stated hypothesis are mine, and I take full responsibility for the accuracy of its contents. The decay curves in Figure 4 are illustrative of a hypothesis and are not data.